1378
doi:10.1017/S1431927619007621
Microsc. Microanal. 25 (Suppl 2), 2019

Practices in Data Management to Significantly Reduce Costs in Cryo-EM

Mario J. Borgnia[1,2*] and Alberto Bartesaghi[2]

[1.] Genome Integrity and Structural Biology Laboratory, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, NC, USA.
[2.] Department of Biochemistry, Duke University School of Medicine, Durham, NC, USA.
* Corresponding author: mborgnia@nih.gov

Structure determination by cryo-electron microscopy (cryo-EM) relies on imaging a specimen consisting of macromolecular complexes embedded in a thin layer of glassy ice. Data collected at the microscope consists of thousands of two-dimensional projection images (micrographs) each depicting tens to hundreds of individual macromolecules. Frozen hydrated biological specimens are composed of weakly scattering elements and because of their sensitivity to radiation need to be imaged under low electron doses. As a consequence, the signal to noise ratio (SNR) of the images is extremely low. To overcome this limitation, thousands of two-dimensional projection images of the complex of interest are computationally aligned and combined to obtain a three-dimensional density map.

The introduction of direct detection devices (DDDs) featuring high detective quantum efficiency marked the beginning of the "resolution revolution" by enabling the determination of atomic resolution structures of macromolecular complexes by cryo-EM. Unlike previous cameras, DDDs record images as a series of frames at a fast rate that makes them capable of counting incident electrons. Images collected in this manner can be saved as "movies" or single micrographs calculated by summing up frames. Preservation of frame data requires significantly more storage space but allows for the correction of several distortions. In addition to drift caused by mechanical and thermal instability, it has been shown that objects imaged in a cryo-electron microscope experience beam induced movement[1]. Images recorded in movie mode can be corrected for movement by tracking the objects throughout the movie prior to summing the frames[2][1]. The effects of radiation are resolution dependent, high resolution features fade earlier during exposure. The SNR of the final micrograph can therefore be increased by weighting the contribution of different components along the "resolution" (frequency) axis of the frames prior to summation. Optical distortions can also be modelled and corrected in movies. The methodology employed in these corrections, critically important for solving structures at atomic resolution[3], is rapidly evolving. Thus, it is highly advisable to preserve as much of the original data as possible. These image files, often consisting of 60 frames each containing 24 megapixels are larger than 5 Gb.

Cryo-EM specimens are typically supported by a metal (copper, gold, etc) grid coated by thin (10-20 nm) substrate (carbon, gold, etc.) that perforated in a pattern with micrometer sized circular holes. The specimen is embedded in a thin layer of vitreous ice inside each hole. Data collection procedures involve detecting a hole, centering the stage (mechanically) on the target, adjusting the imaging conditions in an adjacent area of carbon and recording the image. Several automated data collection software packages are available both from academic (Serial EM, Leginon, UCSF-Image, etc) and commercial sources (Gatan Latitude, FEI EPU, etc). The throughput of data collection is largely limited by two factors: the speed of the detector and the time that it takes to move from target to target. Electron counting relies on reducing the probability of more than one electron hitting each pixel on a given frame. The optimal dose rate for the detector in counting mode depends on its frame rate. Thus, whereas a FEI Falcon 3ec detector requires

a dose rate of 0.8 e^-/pixel/s, the figure goes up 4-8 for the Gatan K2[4] and ~15 for the K3. This establishes a hard limit for the throughput of data collection. Recording a total dose of 50 e^-/Å on an image with a pixel size of 1 Å at the specimen level will take ~60 s on a Falcon 3ec, 12 s for the K2 and 3s for the K3. After considering the time it takes to write the image to disk the maximum throughput is 60, 240 and 720 image per hour respectively. Stage movement and settling; and adjustment of imaging parameters takes, in aggregate, close to 60 seconds, effectively limiting throughput to 30, 48 and 55 images per hour, respectively. Novel approaches to data collection in advanced instruments that involve the use of calibrated beam tilt and image shift to carefully control targeting over a large number of holes can dramatically increase throughput with faster detectors. At the Duke Krios, we routinely collect over 240 images per hour on a FEI Titan Krios equipped with a Gatan K3 detector by imaging nine adjacent holes for each stage movement.

A typical dataset for single particle cryo-EM consisting of 4000 images can now be collected in a single day, posing a series of challenges for a facility. At an average of 4 GB per movie, this dataset would occupy a total of 16 TB of storage space. Transfer of each file over a 1Gb ethernet connection would take over 30 seconds, slower than the rate of data collection. The total volume of data collected over a year of operation would amount to over three petabytes.

Good practices can significantly reduce cost of big data in Cryo-EM. These include optimization of data storage, data reduction at the source, and transfer; and quality control by in-line pre-processing. The sparsity of the signal in electron counting mode resulst in a small number of electrons per pixel per frame. This small integer number can be represented with less than 4 bits. Because the sensitivity of different pixels on a camera is affected by the physical properties of the detector, the relative signals of the pixels need to be normalized multiplying by a gain reference image composed of real numbers. This brings the precision of each pixel to a real value that needs to be represented by 32 bits. The gain reference remains constant over the course of a session and can therefore be extracted from the movies reducing the size of movies by a factor of 4, to 1 GB in our example. Local file transfer then peaks at 450 images per hour over a 1 Gb network link. While a matrix of real numbers is typically compressed by a factor of 2, integer matrices are compressed by factors of 5-10, reducing file size to a much more manageable 200 MB reducing data storage requirements to less than 200 TB per year. Data compression and transfer can be accelerated using software packages that make use of multi-core processors. Several open source software packages that perform in-line image processing give the opportunity of reducing the volume of data by eliminating bad data at the source. These packages provide real-time parameters of data collection that help correct faults. Widely available academic automated data transfer systems make it possible to efficiently distribute data to collaborators and upload it to the cloud. Finally, a number of image processing packages that have been adapted and optimized to be executed in the cloud reduce the cost of image processing by enabling the use of professionally managed computer resources on demand [5].

References:
[1] AF Brilot et al., J Struct Biol. **177** (2012), p. 630.
[2] X Li et al., Nat Methods. **10** (2013), p. 584.
[3] A Bartesaghi et al., Structure. **26** (2018), p. 848.
[4] P-L Chiu et al., J Struct Biol. **192** (2015), p. 163.
[5] The authors acknowledge funding from the US National Institutes of Health Intramural Research Program; US National Institute of Environmental Health Sciences (NIEHS) (ZIC ES103326 to Mario J. Borgnia).